Introduction to Statistical Inference

Sina Pirmoradian, Amirali Soltani Tehrani
sinapirmoradian@ut.ac.ir, aa.soltanitehrani@gmail.com
Instructor: Mohammad-Reza A. Dehaqani
Deadline:
23 Dey 1402

I. Introduction

From examining the variance in family expenditures to exploring the subtleties of binomial distributions, each problem is designed to challenge your conceptual grasp and technical proficiency. The dichotomy of parametric versus non-parametric tests will encourage you to consider the robustness and applicability of different statistical tools. Moreover, the incorporation of simulation techniques will provide you with a hands-on experience of data analysis, reinforcing theoretical knowledge with practical application.

As you progress, you will encounter datasets that require you to employ statistical tests to uncover underlying patterns and relationships. This will not only test your ability to execute statistical procedures but also your capacity to interpret and draw meaningful conclusions from data. The problems are carefully chosen to reflect the complexity and variety of challenges statisticians face in the tech-driven world today.

Whether you're exploring the independence of categorical data, assessing normality, or evaluating the power of a test, this assignment is structured to build your competence in statistical reasoning and decision-making. By the end of this assignment, you should be able to confidently navigate through datasets, apply appropriate statistical tests, and effectively communicate your findings.

Remember, the goal is not just to solve problems but to cultivate a statistical mindset that will serve you throughout your career. Approach each question with curiosity, diligence, and an open mind. Good luck!

II. Non-Parametric Tests

Problem 1

The starting salaries of 17 software developers are 43, 47, 52, 68, 72, 55, 61, 44, 58, 63, 54, 59, 77, 36, 80, 53, 60 thousands of dollars.

1) Does the 5% level Wilcoxon signed rank test provide significant evidence that the median starting salary of software developers is above $50,000? Explain.
2) Use a Mann-Whitney-Wilcoxon Rank Sum Test to answer this question.

Problem 2

Suppose we wanted to know if people's ability to report words accurately was affected by which ear they heard them in. To investigate this, we performed a dichotic listening task. Each participant heard a series of words, presented randomly to either their left or right ear, and reported the words if they could. Each participant thus provided two scores: the number of words that they reported correctly from their left ear, and the number reported correctly from their right ear.

Table 2.1 Number of Words Reported

Participant	Left Ear	Right Ear
1	25	32
2	29	30
3	10	7
4	31	36
5	27	20
6	24	32
7	27	26
8	29	33
9	30	32
10	32	32
11	20	30
12	5	32

1) Do participants report more words from one ear than the other?
2) Use a Mann-Whitney-Wilcoxon Rank Sum Test to answer this question.

Problem 3

Fans Imagine we want to analyze the age of male and female fans of a small new band. We want to test the hypothesis whether male fans are older than female fans. We collected a sample of 50 male fans and 50 female fans. Now, we want to analyze the age distribution to determine if there is a significant age difference between male and female fans.

Men's Age: 52, 18, 27, 12, 24, 17, 68, 25, 12, 9, 51, 44, 42, 34, 44, 15, 21, 66, 61, 32, 31, 20, 6, 13, 34, 38, 45, 17, 16, 15, 36, 21, 29, 21, 29, 9, 33, 15, 37, 27, 31, 15, 57, 37, 27, 31, 38, 27, 60, 23

Women's Age: 36, 49, 20, 31, 51, 31, 15, 16, 39, 70, 52, 16, 39, 34, 18, 34, 30, 18, 26, 18, 25, 16, 39, 49, 22, 37, 39, 21, 16, 63, 45, 43, 17, 28, 29, 23, 42, 23, 28, 55, 41, 18, 23, 8, 13, 26, 13, 27, 28, 18

1) Make a histogram of the original data for the ages of men and women.
2) Show whether the distribution is normal or not using statistical methods.
3) Can we use parametric tests in this problem? Examine all of the assumptions.
4) Can we transform the data to the normal distribution? Implement it and recheck the distribution.
5) If the parametric requirements are met, perform a parametric test and analyze the results.
6) Perform a non-parametric test on the original data and compare the results with the parametric one. Is there any difference? Compare the powers of two tests.
7) Which test is more appropriate in this situation? Why?

Problem 4

Consider a two-way contingency table with three rows and three columns. Suppose that, for $i = 1, 2, 3$ and $j = 1, 2, 3$, the probability p_{ij} that an individual selected at random from a given population will be classified in the ith row and the jth column of Table 4.1.

Table 1 Data for Question 4

0.15	0.09	0.06
0.15	0.09	0.06
0.20	0.12	0.08

$$H_0 : p_{ij} = p_i + p_j \text{ for } i = 1, \ldots, R \text{ and } j = 1, \ldots, C,$$
$$H_1 : \text{The hypothesis } H_0 \text{ is not true.} \quad (4.1)$$

1) Show that the rows and columns of this table are independent by verifying that the values p_{ij} satisfy the null hypothesis H_0 in Eq. (4.1).
2) Generate a random sample of 300 observations from the given population using a uniform pseudo-random number generator. Select 300 pseudo-random numbers between 0 and 1 and proceed as follows: Since $p_{11} = 0.15$, classify a pseudo-random number x in the first cell if $x < 0.15$. Since $p_{11}+p_{12} = 0.24$, classify a pseudo-random number x in the second cell if $0.15 \le x < 0.24$. Continue in this way for all nine cells. For example, since the sum of all probabilities except p_{33} is 0.92, a pseudo-random number x will be classified in the lower-right cell of the table if $x \ge 0.92$.
3) Consider the 3×3 table of observed values N_{ij} generated in part (b). Pretend that the probabilities p_{ij} were unknown, and test the hypotheses (4.1).
4) (Optional) If all the students in a class carry out Question 4 independently of each other and use different pseudo-random numbers, then the different values of the statistic Q obtained by the different students should form a random sample from the χ^2 distribution with four degrees of freedom. If the values of Q for all the students in the class are available to you, test the hypothesis that these values form such a random sample.

Problem 5

Suppose that an experiment is carried out to see if there is any relation between a man's age and whether he wears a mustache. Suppose that 100 men, 18 years of age or older, are selected at random, and each man is classified according to whether or not he is between 18 and 30 years of age and also according to whether or not he wears a mustache. The observed numbers are given in Table 5.1.

Table 5.1 Data for Question 5

	No moustache	Wears a moustache
Between 18 and 30	28	12
Over 30	52	8

1) Test the hypothesis that there is no relationship between a man's age and whether he wears a mustache.

Problem 6

Suppose that 300 persons are selected at random from a large population, and each person in the sample is classified according to blood type, O, A, B, or AB, and also according to Rh, positive or negative. The observed numbers are given in Table 6.1.

Table 6.1 Data for Question 6

	O	A	B	AB
Rh positive	82	89	54	19
Rh negative	13	27	7	9

1) Test the hypothesis that the two classifications of blood types are independent.

Problem 7

Suppose that the ordered values in a random sample of five observations are $y_1 < y_2 < y_3 < y_4 < y_5$. Let $F_n(x)$ denote the sample c.d.f. constructed from these values, let $F(x)$ be a continuous c.d.f., and let D_n be defined by Eq. (7.1).

Let F_n be the sample c.d.f. from an i.i.d. sample $X_1, \ldots, X_n$ from the c.d.f. F. Define:

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)| \quad (7.1)$$

1) Prove that the minimum possible value of D_n is 0.1, and prove that $D_n = 0.1$ if and only if $F(y_1) = 0.1$, $F(y_2) = 0.3$, $F(y_3) = 0.5$, $F(y_4) = 0.7$, and $F(y_5) = 0.9$.
2) Under the conditions of the question, prove that $D_n \le 0.2$ if and only if $F(y_1) \le 0.2 \le F(y_2) \le 0.4 \le F(y_3) \le 0.6 \le F(y_4) \le 0.8 \le F(y_5)$.

Problem 8

Consider the dataset provided in Table 8.1, comprising 25 values. Investigate their distributional characteristics through different tests. Utilize statistical tests to explore the nature of the given dataset, aiming to determine its origin or distribution.

1) **Uniform Distribution Hypothesis**
   Use the Kolmogorov-Smirnov test to assess whether the 25 values in Table 8.1 form a random sample from the uniform distribution on the interval [0, 1].

Table 8.1 Data for Question 8

0.42	0.06	0.88	0.40	0.90
0.38	0.78	0.71	0.57	0.66
0.48	0.35	0.16	0.22	0.08
0.11	0.29	0.79	0.75	0.82
0.30	0.23	0.01	0.41	0.09

2) **Continuous Distribution Hypothesis**
   Apply the Kolmogorov-Smirnov test to investigate if the 25 values from Question 8 constitute a random sample from the continuous distribution described by the probability density function (p.d.f.) $f(x)$.

$$f(x) = \begin{cases} \frac{3}{2} & \text{for } 0 < x \leq 1 \\ \frac{1}{2} & \text{for } \frac{1}{2} < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

3) **Posterior Probability Assessment**
   Considering the prior probabilities related to Question and 2 conditions, determine the posterior probability that the 25 values in Table 8.1 were obtained from a uniform distribution given the prior assumptions.

Problem 9

In this Question we explore the relationship between two unknown distributions, $F(x)$ and $G(x)$, using collected data samples.

1) **Comparison of Two Distributions**
   Given 25 observations from $F(x)$ and 20 observations from $G(x)$, conduct the Kolmogorov-Smirnov test to verify if $F(x)$ and $G(x)$ are identical functions.

Table 9.1 First sample for Question 9

0.61	0.29	0.06	0.59	-1.73
-0.74	0.51	-0.56	-0.39	1.64
0.05	-0.06	0.64	-0.82	0.31
1.77	1.09	-1.28	2.36	1.31
1.05	-0.32	-0.40	1.06	-2.47

Table 9.2 Second sample for Question 9

2.20	1.66	1.38	0.20
0.36	0.00	0.96	1.56
0.44	1.50	-0.30	0.66
2.31	3.29	-0.27	-0.37
0.38	0.70	0.52	-0.71

2) **Shifted Distribution Comparison**
   Revisit the initial conditions and examine the hypothesis that random variables X and Y ($X + 2$ and Y) have equivalent distributions, utilizing the Kolmogorov-Smirnov test.
3) **Scaled Distribution Comparison**
   Building upon the previous analyses, utilize the Kolmogorov-Smirnov test to test the hypothesis that random variables X and Y ($3Y$ and X) share the same distribution, considering the conditions from earlier questions.

Problem 10

Using the Titanic dataset, create a frequency table to display the association between "sex" and "survive" variables.

1) Part 1: Creating Frequency Table and Visualization.
   a) Read the Titanic dataset from the CSV file.
   b) Create a contingency table for the variables "sex" and "survive".
   c) Save this table as `sex_survive_table`.
   d) Visualize the association between these variables using a mosaic plot.
   e) Display the contingency table showing the count of individuals based on gender and survival status.
   f) Generate a mosaic plot visualizing the relationship between gender and survival, specifying colors and axis labels for better interpretation.
2) Part 2: χ^2 Contingency Test and Fisher's Exact Test.
   a) Conduct the χ^2 test on the `sex_survive_table` contingency table.
   b) Verify the assumptions for the χ^2 test.
   c) (Optional) Execute Fisher's exact test for the same contingency table.
   d) Display and analyze the results of the χ^2 test, including the chi-squared value, degrees of freedom, and the p-value indicating the association significance.
   e) (Optional) Present the Fisher's exact test results, including the p-value, confirming the association between gender and survival without approximations.

III. Parametric Tests

Problem 11

Short-answer questions:

1) True or false, and explain briefly.
   a) Even a highly significant difference might be the result of chance.
   b) A large, noteworthy number is one that is statistically significant.
   c) When a p-value is 4.7% as opposed to 5.2
2) Which of the following questions does a test of significance deal with?
   a) Is the difference due to chance?

b) Is the difference important?
c) What does the difference prove?
d) Was the experiment properly designed?

3) Regarding box X, the null hypothesis being tested by two investigators is that its average is equal to 50. Regarding the alternative hypothesis, they both concur that the average is not equal to 50. They concur to use a two-tailed z-test as well. Using replacements, the first investigator draws 100 tickets at random from the box. With replacement, the second investigator draws 900 tickets at random. For both investigators, the SD is 10. True or false: the investigator with the lesser p-value is the one whose average is furthest away from 50. Give a brief explanation.
4) Assume that if the z-test indicates that there is a "statistically significant" difference between the percentage of White employees at a company and the percentage of White employees in the surrounding geographic area, then courts have found a prima facie case of discrimination against the company. Let's say that 10% of the population in a certain city is white. Assume moreover that every company in the city employs workers using a procedure that is essentially random sampling in terms of race. Would the z-test ever find any of these companies to be discriminating? Give a brief explanation.

Problem 12

Assume that you apply the following decision rule to examine the premise that a coin is fair: If the number of heads you witness in a single sample of 100 tosses falls between 40 and 60, you accept the null hypothesis that the coin is fair; if not, you reject the null hypothesis.

1) What is the likelihood that the hypothesis will be rejected even though it is true?
2) For this decision rule, what is the equivalent α?
3) (**Programming Part**): Draw, using just one drawing, the binomial distribution of the number of heads in 100 tosses for p = 0.5 and p = 0.7. Mark α and β in this drawing. Given that p(heads) = 0.7 indicates that the coin is unfair, what is the likelihood of adopting the null hypothesis? What is the strength of the test in first parts to identify that this coin is unfair if the real chance of receiving a head is p=0.7? Repeat this part with more number of iterations in a reasonable way for yourself and explain it!

Problem 13

A random sample of fifty public high school students and a sample of sixty-five Catholic high school students were given a test. The average grade in the public school sample was 70, but the Catholic high school had a mean grade of 74. It is well known that test results at both schools are dispersed normally. At the public school, the standard deviation of the results was 4, whereas at the Catholic school, it was 6.

1) Which statistical test, and for what reason, is suitable to determine if the mean scores at the two schools differ significantly?
2) Test whether the difference is significant, with $\alpha = 0.05$.

Problem 14

The idea that people who play video games have superior spatial perceptions than people who don't is something that a researcher want to investigate. Twenty pupils who play video games less than an hour a week and fifteen who play at least ten hours a week are given a spatial aptitude exam by him. The average score for those who play video games is 120 for spatial ability. For the non-players, the average spatial aptitude score is 100. Twenty and fifty are the standard deviations.

1) Is this an observational study or a controlled experiment? (optional)
2) Is the difference in mean aptitude score for the two groups significantly different at the 0.05 level? At the 0.01 level?
3) Would a substantial difference seen in this study indicate that playing video games enhances spatial aptitude? What other reason may there be?

Problem 15

When testing a hypothesis on the mean of a Gaussian distribution with known variance, the power of a one-sided alternative is

$$\Phi\left(z_\alpha + \frac{n^{\frac{1}{2}}(\mu_A - \mu_0)}{\sigma}\right)$$

1) How strong is a research if its sample of 49 participants comes from a Gaussian distribution with a variance of 49 and an anticipated difference between μ_0 and μ_A of 0.5? A type I error of 0.05 is required.
2) What is the type II error?
3) Find a sample size so that the power for this study is 0.99.

Problem 16

Suppose $X \sim \text{Binomial}(100, p)$ and consider the following test:

$$H_0 : p = 0.5$$
$$H_1 : p \neq 0.5$$

If $|X - 50| > 10$, then H_0 is rejected. Answer the following questions:

1) What is the value of α?
2) Draw an approximate power diagram as a function of p.

Problem 17

For a test of hypothesis about the mean of a population with an exponential distribution, where the likelihood function is given by:

$$f(x|\theta) = \theta e^{-\theta x}$$

and $X_1, ..., X_n$ is a random sample, the likelihood ratio test is used to decide between the following hypotheses:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Problem 18

In a test for the average weight of newborns in a hospital, the following data are obtained from a sample with 50 sample size: $\bar{x} = 25.9$, $s = 5.6$. The hypotheses are:

$$H_0 : \mu \geq 28$$
$$H_a : \mu < 28$$

1) Assuming $\alpha = 0.5$, what conclusions can you draw from this sample?
2) If the true value of μ is not greater than 27, find the probability of committing a Type II error.

Problem 19

In a study to estimate the average duration of a certain type of surgical procedure, the duration times for 40 surgeries are recorded. The sample mean is 130 minutes, and the standard deviation is 5 minutes. The 95% confidence interval for the mean duration of the surgery, which is constructed using the bootstrap method with 1000 replications, is between 128.5 minutes and 131.5 minutes.

1) What conclusions can be drawn from this study?
2) How can the bootstrap method be used to construct a confidence interval for the median?
3) What is the advantage of using the bootstrap method over parametric methods in this study?
4) Is the bootstrap method always the best method for estimating parameters? Explain.
5) What are bootstrap statistics and how are they calculated?

Problem 20

Programming Question: Generate a sample of size 50 from a beta distribution with parameters $a = 2$ and $b = 5$.

1) Consider the median of the sample above as M. Test the following hypothesis with a nonparametric test.

$$H_0 : M = 0.4$$
$$H_1 : M > 0.4$$

2) Calculate the power for the above test.
3) Increase the value of 0.4 in the first part to 0.6 and repeat both parts.
4) Compare the results obtained in part c with part a and justify the change in the power.

Problem 21

Programming Question: Suppose we have:

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(0, \sigma_x^2)$$
$$Y_1, Y_2, \ldots, Y_n \sim \mathcal{N}(0, \sigma_y^2)$$

1) What method do you suggest for testing the hypothesis $H_0 : \sigma_x = \sigma_y$?
2) Using samples from two normal distributions with variances 3 and 10, examine the above method.

Problem 22

Programming Question: Table 1. shows the number of sons in families that have 12 children. Assuming that X_i is the random variable corresponding to the number of sons in these 6115 families,

Number of sons	Frequency
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3

a) Introduce statistics that can be used to test the following hypothesis:

$$H_0 : X_1, X_2, \ldots, X_{6115} \sim \text{Binomial}(12, 0.5)$$

b) Using simulation, obtain the distribution of the introduced statistic and draw its histogram.
c) Based on the obtained distribution, calculate the p-value for the statistic calculated in part a. With $\alpha = 0.05$, is H_0 rejected?

IV. Submission

For the programming section, each student is required to submit a well-structured, typed PDF report that presents a concise summary of their analysis. The report should include the figures mentioned in the problem description and offer a detailed discussion of each. Please avoid uploading theoritical problem in .jpg format and upload them in a single .pdf file.

For each section of the report, a separate script is expected, which can be written in MATLAB (.m), Python 3 (.py or .py3), or R (.r). Avoid submitting scripts in formats like MATLAB live scripts, Python notebooks, or R Markdown. It is crucial that the submitted code is compatible with the grader's system. Be sure to include all relevant functions and any non-standard libraries used in your code.

The report should be treated as an academic piece of writing, and it should not contain any code snippets or explanations of coding logic. Instead, it should provide the author's insights about the results and demonstrate a strong grasp of the reference article. Academic reports typically maintain a concise and highly formal tone.

Each section of the report should briefly outline the hypothesis being tested. The responsibility for designing and implementing the tests lies with the students, as does explaining the results. Interpretations should be comprehensive without unnecessary verbosity.

The report can be written in either Persian or English, with no preference for either. In Persian reports, use B Nazanin

with a font size of 14 for the text body and B Titr with a font size of 18 for titles. English reports should use Times New Roman 12 for the body text and Times New Roman 16 for titles. Sentences should be written in the passive tense. In Persian reports, the correct usage of the zero-width non-joiner is mandatory. In all reports, equations, figures, and tables must be labeled with unique numbers and referenced accordingly. Referring to figures as ”the following figure,” ”the figure above,” and similar expressions is considered incorrect.

Every figure in the report should be accompanied by a descriptive caption below it, while tables should have captions above them. Feel free to use footnotes and citations as necessary for clarity and proper attribution.